Homework#1 (Total score = 100)

COMPSCI- 5590-0021 Econometrics of Data Science
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF MISSOURI-KANSAS CITY

Due on Tuesday, March 4, 2025 at 11:59 PM (in Canvas)

"It must be an individual submission. Any kind of copying or corroboration will be severely penalized".

This dataset "ms-drg-2018.csv" contains data about the Medicare Severity Diagnosis-Related Groups for the year 2018. It includes data on medical facility names, IDs, street addresses, names of cities and states where the facilities are located, ownership type, and DRG definitions and codes of certain diseases. The dataset also contains the total number of patients discharged, average covered charges, average total payments, and average Medicare payments associated with a specific DRG or patient category under each medical facility. Here, the "average covered charge" denotes a measure of the average amount hospitals bill for services within a specific DRG or patients' disease category. The variable "average total payments" represents the average reimbursement a hospital receives (from all sources) for treating patients in a specific DRG or category. The "average Medicare payment" refers to the average reimbursement a hospital receives from Medicare for treating patients (a subset of the total payment) in a specific DRG or disease category.

We are interested in conducting some descriptive analysis using this medicare dataset. You must use **R**-studio to solve the following questions:

1. Generate summary statistics (including skewness and kurtosis) of the total number of patients discharged, average covered charges, average total payments, and average Medicare payments for Missouri and California states. Interpret and compare the results. (20)

2. Generate box plots of the "average total payments" for Missouri and California, considering only those observations that fall under a 2-standard deviation of their corresponding means. Compare the plots in terms of median, spread, skewness, and outliers (if any). (20)

[Hints: for each state, first select those observations (rows) that fall under a 2-standard deviation of the mean, then store them in a separate data frame and draw boxplots].

3. Generate the summary statistics, including minimum, maximum, and three-quartile values of "average covered charges" for the three urban states (California, New York, and Florida). Also, generate the summary statistics of the same variable for the three urban-rural mixed states (New Mexico, North Dakota, and Wyoming). Compare the results between urban states and urban-rural mixed states. (20)

4. Generate a data frame considering only those observations (rows) that are related to any kind of "CARDIAC" disease. Then, calculate the summary statistics, including average total discharges, median total discharges, average covered charges, and average total payments, for the following six states: California, Wyoming, Idaho, New York, Kansas, and Missouri. Comment on the summaries across six states. (30)

[Hints: you can find a cardiac-related list of diseases under the variable drg_definition. Note that 32 unique categories of cardiac-related disease are reported in the variable drg_definition, and 2,018 observations contain any of these 32 categories for the given six states. Use the 'ddply' command to find the summary statistics of the given items grouped by the six states.]

5. Generate a bar plot, placing six states, California, Wyoming, Idaho, New York, Kansas, and Missouri, on the x-axis and the average number of total discharges related to cardiac-related diseases on the y-axis. Comment on the plot. (10)

[Hint: Use the data frame and summary statistics you generated in question 4 to generate this bar plot].

Submission guidelines:

- In the document, write down a question and then write your answer afterward.
- In the answer script, write your name and student ID on the right-hand side of the header.
- Please type your response. No handwritten submission will be accepted.
- Upload a pdf file containing the results, graphs, reasoning, and interpretations.
- Upload the R file.
- Please write the file name (both pdf and R) in this format: COMPSCI-5590-hw1-your last name